Behavior based adaptive call predictor

Behavior-Based Adaptive Call Predictor

SANTI PHITHAKKITNUKOON, Newcastle University
RAM DANTU, University of North Texas
ROB CLAXTON, British Telecommunications
NATHAN EAGLE, Massachusetts Institute of Technology

Predicting future calls can be the next advanced feature of the next-generation telecommunication networks as the service providers are looking to offer new services to their customers. Call prediction can be useful to many applications such as planning daily schedules, avoiding unwanted communications (e.g. voice spam), and resource planning in call centers. Predicting calls is a very challenging task. We believe that this is an emerging area of research in ambient intelligence where the electronic devices are sensitive and responsive to people's needs and behavior. In particular, we believe that the results of this research will lead to higher productivity and quality of life. In this article, we present a Call Predictor (CP) that offers two new advanced features for the next-generation phones namely "Incoming Call Forecast" and "Intelligent Address Book." For the Incoming Call Forecast, the CP makes the next-24-hour incoming call prediction based on recent caller's behavior and reciprocity. For the Intelligent Address Book, the CP generates a list of most likely contacts/numbers to be dialed at any given time based on the user's behavior and reciprocity. The CP consists of two major components: Probability Estimator (PE) and Trend Detector (TD). The PE computes the probability of receiving/initiating a call based on the caller/user's calling behavior and reciprocity. We show that the recent trend of the caller/user's calling pattern has higher correlation to the future pattern than the pattern derived from the entire historical data. The TD detects the recent trend of the caller/user's calling pattern and computes the adequacy of historical data in terms of reversed time (time that runs towards the past) based on a trace distance. The recent behavior detection mechanism allows CP to adapt its computation in response to the new calling behaviors. Therefore, CP is adaptive to the recent behavior. For our analysis, we use the real-life call logs of 94 mobile phone users over nine months, which were collected by the Reality Mining Project group at MIT. The performance of the CP is validated for two months based on seven months of training data. The experimental results show that the CP performs reasonably well as an incoming call predictor (Incoming Call Forecast) with false positive rate of 8%, false negative rate of 1%, and error rate of 9%, and as an outgoing call predictor (Intelligent Address Book) with the accuracy of 70% when the list has five entries. The functionality of the CP can be useful in assisting its user in carrying out everyday life activities such as scheduling daily plans by using the Incoming Call Forecast, and saving time from searching for the phone number in a typically lengthy contact book by using the Intelligent Address Book. Furthermore, we describe other useful applications of CP besides its own aforementioned features including Call Firewall and Call Reminder.

This work is supported by the National Science Foundation under grants CNS-0627754 (Detecting Spam in IP Multimedia Communication Services), CNS-0619871 (Development of a Flexible Platform for Experimental Research in Secure IP Multimedia Communication Services), and CNS-0551694 (A Testbed for Research and Development of Secure IP Multimedia Communication Services). Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Authors' addresses: S. Phithakkitnukoon (corresponding author), Culture Lab, Newcastle University, UK; email: santi@newcastle.ac.uk; R. Dantu, Department of Computer Science and Engineering, University of North Texas, Denton, TX; R. Claxton, BT Group plc, Ipswich IP5 3RE, UK; N. Eagle, Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

DOI 10.1145/2019583.2019588 http://doi.acm.org/10.1145/2019583.2019588

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*; J.4 [**Social and Behavioral Sciences**]: Sociology; I.5.2 [**Pattern Recognition**]: Design Methodology—*Pattern analysis*

General Terms: Human Factors

Additional Key Words and Phrases: Prediction, behavior, call logs, call matrix, trace distance, convergence time

ACM Reference Format:
Phithakkitnukoon, S., Dantu, R., Claxton, R., and Eagle, N. 2011. Behavior-based adaptive call predictor. ACM Trans. Autonom. Adapt. Syst. 6, 3, Article 21 (September 2011), 28 pages.
DOI = 10.1145/2019583.2019588 http://doi.acm.org/10.1145/2019583.2019588

1. INTRODUCTION

With the rapid development of telecommunication technologies and the fast-growing number of users on the networks, more and more services are being offered by telephone service providers. As phone companies are engaged in a fiercely fought race to add new advanced features to their products, today's phones become more than just communication devices. In addition to the voice calls, they can be used to send email, take photos and videos, navigate the Internet, play online games, listen to music, and even conduct bank transactions. Mobile phones have become an indispensable part of life for many people who perform several activities using their phones such as reading novels and books on mobile phones [Web Japan 2006], shopping by mobile phones [Trendhunter 2006], and getting cyber counselors for quitting smoking [Ntuli 2007]. In 2005, Google filed a patent including details about the new Google Phone (GPhone) that could predict what a user is searching for or the words that are typed in a text message by taking into account the user's location, previous searching/messaging history, and time of the day.

However, none of these aforementioned features offers the ability to predict future calls. To the best of our knowledge, no scientific research has been reported in predicting the incoming/outgoing calls for phone services. Predicting calls using just the call history is a challenging task. We believe that this is an emerging area of research in ambient intelligence where the electronic devices are sensitive and responsive to people's needs and behavior.

Prediction plays an important role in many applications and it is widely applied in various areas such as weather, economic, stock, disaster (e.g., earthquake and flooding), network traffic, and call center forecasting. Companies use predictions of demands for making investments and allocating resources efficiently. Call centers use predictions of workload to prepare the right number of staff in place to handle it. Network administrators use traffic predictions to assess future network capacity requirements and to plan network development so as to better use network resources and to provide better quality of services. Prediction is also applied in the human behavior study by combining the computer technology and social networks, for example, Eagle and Pentland [2005, 2006] and Eagle et al. [2007].

Predicting the next-day incoming calls can be very useful for scheduling a day (e.g., it can be used to avoid unwanted calls and schedule time for wanted calls). People check weather forecasts before leaving homes and watch for signs of approaching storms to prepare and schedule their days accordingly. Knowing what is coming next gives us supplemental time to think, prepare, and optimize the solutions. We believe that prediction of future calls can be very useful for daily planning and it will become an important element as an initiative decision support for daily life scheduling. People will start the day by checking the weather forecast as well as the call prediction.

Predicting outgoing calls can be used to improve the "last number dialed" functionality that is normally provided on today's phones. Providing the most likely contacts/numbers to be dialed on the top of the list instead of only the recent called contacts/numbers reduces the searching time and enables better life synchronization for the phone user.

In this article, we present a model for predicting incoming and outgoing calls based on the caller/callee and phone user's past communication information. The rest of this article is structured as follows. Section 2 presents the architecture of the Call Predictor (CP). Section 3 describes our real-life datasets. Sections 4 and 5 present the CP's framework, which carries out the receiving/initiating call probability computation along with the discussion on the caller's behavior trend detection. The performance of the CP is then evaluated and discussed in Section 6. Section 7 reemphasizes the autonomous and adaptive characteristics of CP as well as describes applications of CP in the context of the smart phone. The literatures that are related to our work are discussed in Section 8. Section 9 concludes our contribution with a summary and an outlook on future work.

2. CALL PREDICTOR

The Call Predictor (CP) described here is intended to offer the ability to predict incoming calls as well as outgoing calls as the new advanced features to today's personal phones. The CP allows the user to see the "Incoming Call Forecast" of the next 24 hours which may be used as an initiative decision support for the user's daily life scheduling or other purposes. The CP also attempts to provide an improvement over the "last number dialed" functionality that is often provided on phones and communication clients (e.g., VoIP soft-phones). It is common for the user interface on a mobile phone to provide easy access to a list of recently dialed numbers and therefore takes no account of the user's situation (e.g., location, time, social relationship, etc.) to inform a better "guess" of the numbers that the user will find most useful. Therefore, the CP generates a list of the most likely contacts/numbers to be dialed at any given time. This list can then be presented to the user in a number of different ways for different purposes. The principle mode of presentation envisaged is as an "Intelligent Address Book," that is, an address book that anticipates the contacts that the user wants to call and gives these contacts higher precedence in any listing. In this mode, the information presented to the user is not intended as a direct replacement for the normal Address Book functionality (i.e., where the user may search for contacts) but as an improved short-cut device.

The CP makes use of the user's past call logs, that is, incoming/outgoing calls, to build a probabilistic model of calling behavior. The call logs can be derived locally on a single device or from the operator's network (e.g., billing information). Similarly, the model itself can be stored locally on the user's device or maintained "in the network" (in which case it can be shared and made available across a range of devices that the user has access to). The basic architecture of the CP is illustrated in Figure 1.

With the architecture shown in the figure, the CP can operate in two modes upon the user's request: *Incoming Call Forecast* and *Intelligent Address Book*.

(1) *Incoming Call Forecast*: For any time that phone user requests a call prediction of a particular caller, the CP detects the most recent calling trend of the caller and computes the probability of receiving calls of the next 24 hours based on the caller's past history (caller's incoming calls) and the user's call history towards the caller (user's outgoing calls). These historical call logs are maintained by the CP by logging the call-specific information for every call received and made by the user. The computed receiving call probability is then checked with the preconfigured

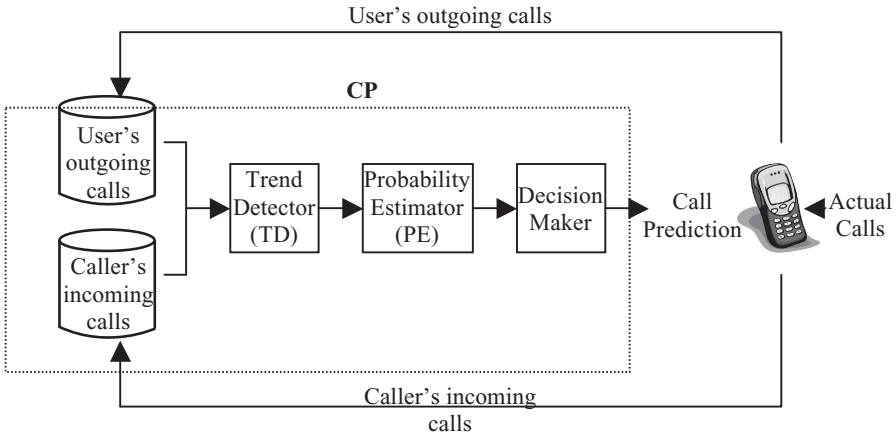

Fig. 1. Basic architecture of Call Predictor (CP).

threshold value to make a decision as to predict "call" or "no call" for each of the next 24-hour time slots.

(2) *Intelligent Address Book*: For any time the phone user attempts to make a call (i.e., unlock the keypad, flip up the phone, etc.), the CP detects the most recent calling trend of the user and computes the probability of making a call to each caller (whom the user previously made calls to) based on user's call history towards the caller (user's outgoing calls) and the caller's past history (caller's incoming calls). The list of the most likely contacts/numbers to be dialed is then generated according to the computed probabilities.

3. REAL-LIFE DATASETS

Everyday phone calls include calls from different sections of our social life. We receive calls from family members, friends, supervisors, neighbors, and strangers. We believe that every person exhibits a unique calling pattern. These calling patterns can be analyzed for predicting the future calls to/from the phone user.

To study calling patterns, we use the real-life call logs of 94 individual mobile phone users over nine months which were collected at Massachusetts Institute of Technology (MIT) by the Reality Mining Project [Massachusetts Institute of Technology 2007]. These 94 individuals are faculty, staff, and students. The datasets include people with different types of calling patterns and call distributions.

We extract 5-tuple information of the call record for each phone user from the datasets: {Date of call, Start time of call, Type of call (Outgoing, Incoming), Call ID, Call duration}. We use our real-life datasets for deriving the traffic profiles for each caller who called the individuals. To derive the profile, we infer the arrival time (time of receiving a call), inter-arrival time (elapsed time between adjacent incoming calls), and inter-arrival/departure time (elapsed time between adjacent incoming and outgoing calls).

4. INCOMING CALL PREDICTION FRAMEWORK

To predict the future incoming calls, a dynamic decision making technique has to be integrated with behavior learning models. These models should incorporate mechanisms for capturing the caller's behavior (based on call arrival time and inter-arrival time), the user's behavior (based on call departure time), reciprocity (based on call inter-arrival/departure time), and caller's behavior trend, to construct the probabilistic

model for the caller's incoming calls and finally generate the "Incoming Call Forecast" (the next-24-hour incoming call prediction.)

4.1. Probability Computation

In our daily life, when we receive a phone call, at the moment of the first phone ring before we look at the caller ID, we often guess who the caller might be. We usually base this estimation on the following information.

—*Caller's past behavior*. Each caller has a unique calling pattern. These patterns can be observed through *historical calling time* (we normally expect a call from a caller who has history of making several calls at some particular time, e.g., your spouse likes to call you while you are driving to work in the morning and in the evening after work, and therefore when your phone rings while you are on the way to work or back home, you are likely to guess that it is a phone call from your spouse), and *periodicity of calls* (we normally expect that a caller who calls periodically will repeat the same pattern, e.g., your best friend calls you at about 2:00 PM every Tuesday, and therefore you expect a call from him/her at about 2:00 PM for the next Tuesday).

—*Reciprocity*. The past communication activities between each caller and the user also establish a unique pattern. These patterns can be observed in forms of the number of the user's outgoing calls per caller's incoming call and call interarrival/departure time.

Therefore, we believe that receiving a call is influenced by the caller's past incoming calls and historical call interaction between caller and phone user. The pattern of the caller's incoming calls can be observed from call arrival time and interarrival time. The pattern of call interaction between caller and phone user can be observed from the number of outgoing calls per incoming call and the interarrival/departure time.

4.1.1. Probability Computation Based on Caller's Behavior. Based on the pattern of the caller's call arrival time, callers can be roughly divided into two groups.

(1) *Single-Hop Callers*. There are callers who tend to make more calls at around one particular time of the day and the number of calls gradually decreases as time of the call deviates from that time (favorite time). Thus, we make a hypothesis that call arrival time has a normal distribution $N(\mu_W, \sigma_{2W})$ where μ_W is the mean and σ_W^2 is the variance of call arrival time which can be calculated by (1) and (2) respectively.

$$\mu_W = \frac{1}{N}\sum_{n=1}^{N} w(n) \tag{1}$$

$$\sigma_W^2 = \frac{1}{N}\sum_{n=1}^{N} (w(n) - \mu_W)^2 \tag{2}$$

Let a random variable W represent the arrival time of phone calls, and let $\{w(1), w(2), w(3), \ldots, w(N)\}$ be a set of observed values of arrival times. Here, N is the total number of calls and $w(n)$ is the n^{th} call arrival time. The estimated probability density (pdf) of W is given by (3) where μ_W and σ_W^2 are mean and variance of W.

$$a(w) = \frac{1}{\sqrt{2\pi\sigma_W^2}} e^{-(w-\mu_W)^2/2\sigma_W^2} \tag{3}$$

Hence the probability of receiving a call from caller k between w^{th} and $(w+1)^{th}$ hour slot is given by (4) where $\mu_{W,k}$ and $\sigma_{W,k}^2$ are the corresponding mean and

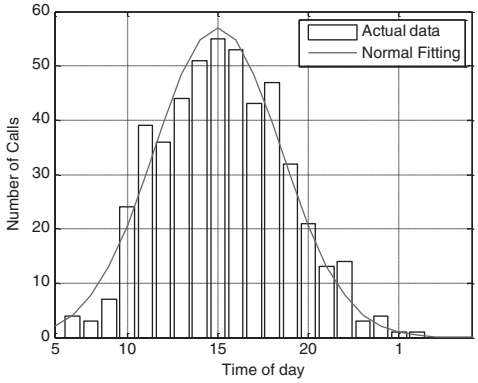

Fig. 2. An example of single-hop caller whose call arrival time is fitted with normal distribution.

variance of call arrival time of caller k and W_k is random variable W of caller k.

$$P_k^a(w) = \Pr\{w \leq W_k < w+1\} = \int_w^{w+1} \frac{1}{\sqrt{2\pi\sigma_{W,k}^2}} e^{-\left(t-\mu_{W,k}\right)^2 / 2\alpha_{W,k}^2} dt \tag{4}$$

To check our hypothesis, we randomly (by visual inspection) select 30 single-hop callers (based on visual inspection) from our datasets and perform the chi-square goodness-of-fit test (or χ^2-test) [Leon-Garcia 1994] which tests the validity of the assumed distribution for a random phenomenon. All 30 single-hop callers pass the test, which is performed using a significant level $\alpha = 0.01$. Note that those callers in our datasets who do not pass the χ^2-test may belong to another group of callers which will be described in the next section.

As an example, in Figure 2 the histogram of the call arrival time over the course of nine months on time-of-the-day scales of a single-hop caller and fitted normal distribution are illustrated where we shift our window of observation to begin at 5AM and end at 4:59AM such that the entire calling pattern is captured in the middle. In fact, we find that it is a proper window of observation for the majority of the callers in our datasets.

(2) *Multihop Callers*. There is another group of callers whose calling patterns based on arrival time are more random. These callers tend to have more than one favorite calling time which draws multiple peaks (or hops) in their arrival time's histograms.

The normal distribution is obviously not suitable for this type of callers. In fact, none of the parametric probability models fits their structures. Therefore, a probability density model must be determined from the data by using nonparametric density estimation. The most popular method is the kernel density estimation which is also known as the Parzen window estimator [Parzen 1962] given by (5).

$$a(w) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{w - w_i}{h}\right) \tag{5}$$

$K(u)$ is kernel function and h is the bandwidth or smoothing parameter. The most widely used kernel is the Gaussian of zero mean and unit variance which is defined by (6).

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \tag{6}$$

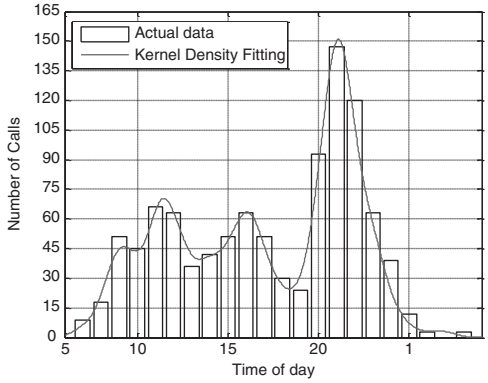

Fig. 3. An example of a multihop caller whose call arrival time is fitted with kernel density estimation.

The choice of the bandwidth h is crucial. Several optimal bandwidth selection techniques have been proposed, for example, Jones et al. [1996] and Wand and Jones [1994]. In this article, we use AMISE optimal bandwidth selection using the Sheather-Jones solve-the-equation plug-in method which was proposed in Sheather and Jones [1991].

Likewise, the probability of receiving a call from caller k between time $w(i)$ and $w(i+1)$ can be calculated similarly to (4) but using the pdf defined in (5). As an example, the histogram of a multihop caller over nine months on time-of-the-day scales and fitted kernel density estimation is illustrated in Figure 3.

We define a call matrix of a single user as a matrix whose entries are call indicators where rows are hours of the day and columns are days of observation. The Call Indicator (CI) indicates if there is at least one incoming call or outgoing call or both incoming call and outgoing call or no call. The CI's values and its indications are given by (7). As an example, a call matrix of 15 days of observation is illustrated in Figure 4.

$$CI = \begin{cases} 0, \text{no call} \\ 1, \text{at least one incoming call} \\ 2, \text{at least one outgoing call} \\ 3, \text{at least one incoming call and one outgoing call} \end{cases} \tag{7}$$

The behavior of the caller can also be observed through the call inter-arrival time. However, inter-arrival time in a normal sense is the elapsed time between temporally adjacent calls made on a per-day basis, which we believe does not accurately represent the caller's behavior, due to human nature. People require a state of rest (represented by night-time hours), and sleeping time causes an inaccuracy in the average inter-arrival time. In fact, it increases that average from its true value. Thus, we believe that the more accurate point of view to observe calling pattern based on inter-arrival time is to scan over each hour time slot of the day through days of observation, that is, capturing the pattern of inter-arrival time by observing each row of the call matrix.

Let a random variable X_i represent inter-arrival of the i^{th} hour slot where $i = 1, 2, 3, \ldots, 24$. A normal distribution $N(\mu_{X,i}, \sigma^2_{X,i})$ is assumed for the call inter-arrival time since no information is available that $\Pr(X_i \leq \mu_{X,i} - c) < \Pr(X_i \geq \mu_{X,i} + c)$ or vice versa therefore it can be safely assumed that $\Pr(X_i \leq \mu_{X,i} - c) = \Pr(X_i \geq \mu_{X,i} + c)$, where the mean $\mu_{X,i}$ and variance $\sigma^2_{X,i}$ of the i^{th} hour can be calculated by (8) and (9),

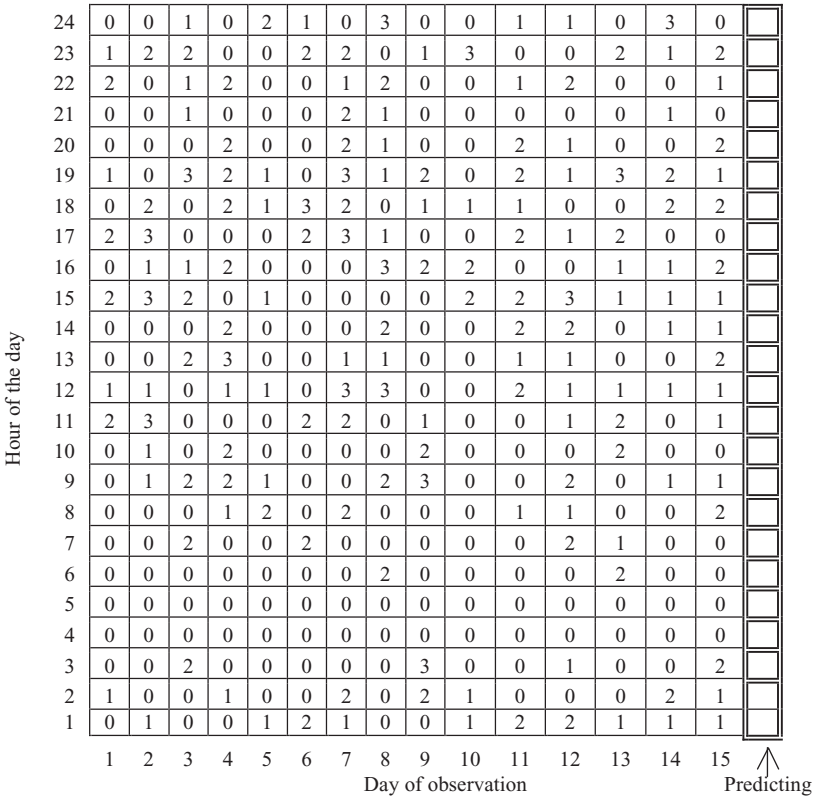

Fig. 4. An example of a call marix of 15 days of observation.

respectively.

$$\mu_{X,i} = \frac{1}{N-1}\sum_{n=1}^{N-1} x_i(n) \tag{8}$$

$$\sigma^2_{X,i} = \frac{1}{N-1}\sum_{n=1}^{N-1} (x_i(n) - \mu_{X,i})^2 \tag{9}$$

The variable N is the total number of calls and $x_i(n)$ is the n^{th} inter-arrival time where both are on the i^{th} hour slot. The inter-arrival time is now treated as a normal random variable X_i that consists of observed values of inter-arrival times $\{x_i(1), x_i(2), x_i(3), \ldots, x_i(N-1)\}$ and its pdf is given by (10).

$$b_i(x_i) = \frac{1}{\sqrt{2\pi\sigma^2_{X,i}}} e^{-(x_i-\mu_{X,i})^2/2\sigma^2_{X,i}} \tag{10}$$

The variable N is the total number of calls and $x_i(n)$ is the n^{th} inter-arrival time where both are on the i^{th} hour slot. The inter-arrival time is now treated as a normal random variable X_i that consists of observed values of inter-arrival times $\{x_i(1), x_i(2), x_i(3), \ldots, x_i(N-1)\}$ and its pdf is given by (10).

For example, if a caller calls on average every 3 days, the chances of receiving a call one day earlier (day 2) or one day later (day 4) are the same.

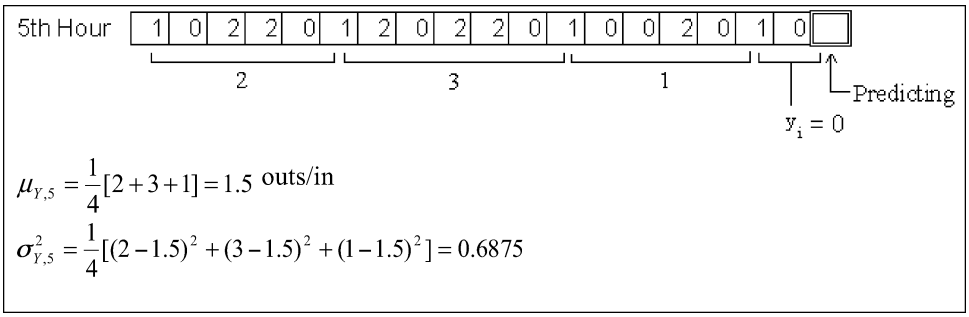

Fig. 5. An example of calculating $\mu_{Y,i}$ and $\sigma^2_{Y,i}$ for one hour slot (5^{th} hour) of 18 days of observation.

The probability of receiving a call during i^{th} hour slot from caller k between x^{th} and $(x+1)^{th}$ day can be calculated by (11).

$$P^b_k(i) = \Pr(x \le X_{i,k} < x+1) = \int_x^{x+1} \frac{1}{\sqrt{2\pi\sigma^2_{X,i,k}}} e^{-(t-\mu_{X,i,k})^2/2\sigma^2_{X,i,k}} dt \tag{11}$$

4.1.2. Probability Computation Based on Reciprocity. As previously mentioned that receiving a call is influenced by not just the caller's behavior but also reciprocity, one way to observe the calling patterns based on reciprocity is to monitor the number of outgoing calls per incoming call. This can give us a good approximation of when the next call can be expected. A normal distribution $N(\mu_{Y,i}, \sigma^2_{Y,i})$ is also assumed for the same reason as in the inter-arrival time case, where the number of outgoing calls per incoming call of the i^{th} hour time slot is represented by a random variable Y_i where the mean $\mu_{Y,i}$ and variance $\sigma^2_{Y,i}$ can be calculated by (12) and (13), respectively.

$$\mu_{Y,i} = \frac{1}{M} \sum_{n=1}^{M-1} y_{Y,i}(n) \tag{12}$$

$$\sigma^2_{Y,i} = \frac{1}{M} \sum_{n=1}^{M-1} (y_{Y,i}(n) - \mu_{Y,i})^2 \tag{13}$$

The variable M is the total number of incoming calls of i^{th} hour and $y_i(n)$ is the number of outgoing calls between the n^{th} and $(n+1)^{th}$ incoming call. The pdf is given by (14).

$$c_i(y_i) = \frac{1}{\sqrt{2\pi\sigma^2_{Y,i}}} e^{-(y_i-\mu_{Y,i})^2/2\sigma^2_{Y,i}} \tag{14}$$

For clarification, an example of computing $\mu_{Y,i}$ and $\sigma^2_{Y,i}$ is illustrated in Figure 5.

Thus, the probability of receiving a call during i^{th} hour slot from caller k between y^{th} and $(y+1)^{th}$ day can be calculated by (15).

$$P^c_k(i) = \Pr(y \le Y_{i,k} < y+1) = \int_y^{y+1} \frac{1}{\sqrt{2\pi\sigma^2_{Y,i,k}}} e^{-(t-\mu_{Y,i,k})^2/2\sigma^2_{Y,i,k}} dt \tag{15}$$

Let a random variable Z_i represent the inter-arrival/departure time of the i^{th} hour. A normal distribution $N(\mu_{Z,i}, \sigma^2_{Z,i})$ is also assumed by the same reason as in the previous

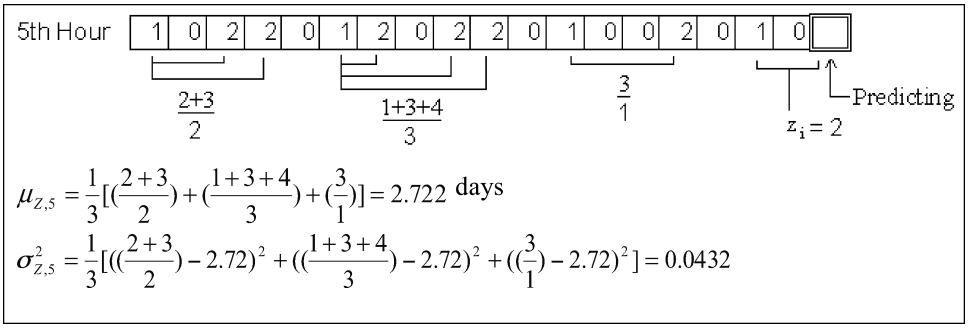

Fig. 6. An example of calculating $\mu_{Z,i}$ and $\sigma^2_{Z,i}$ for one hour slot (5^{th} hour) of 18 days of observation.

cases where $\mu_{Z,i}$ is the mean and $\sigma^2_{Z,i}$ is the variance. These are given by (16) and (17), respectively.

$$\mu_{Z,i} = \frac{1}{L-1}\sum_{n=1}^{L-1} z_i(n) \tag{16}$$

$$\sigma^2_{Z,i} = \frac{1}{L-1}\sum_{n=1}^{L-1} (z_i(n) - \mu_{Z,i})^2 \tag{17}$$

The variable L is the total number of incoming calls of i^{th} hour where $z_i(n)$ is the average inter-arrival/departure time between the n^{th} and $(n+1)^{th}$ incoming calls (an example is illustrated in Figure 6). Thus, the pdf of the inter-arrival/departure time is given by (18).

$$d_i(z_i) = \frac{1}{\sqrt{2\pi\sigma^2_{Z,i}}} e^{-(z_i-\mu_{Z,i})^2/2\sigma^2_{Z,i}} \tag{18}$$

An example is also illustrated in Figure 6 for clarification.

The probability of receiving a call during i^{th} hour slot from caller k between z^{th} and $(z+1)^{th}$ day based on the inter-arrival/departure time can be calculated by (19).

$$P^d_k(i) = \Pr(z \le Z_{i,k} < z+1) = \int_z^{z+1} \frac{1}{\sqrt{2\pi\sigma^2_{Z,i,k}}} e^{-(t-\mu_{Z,i,k})^2/2\sigma^2_{Z,i,k}} dt \tag{19}$$

From (4), (5), (11), (15), and (19), we can infer the probability of receiving a call from "*Caller A*" during i^{th} hour ($P_A(i)$) as the average of the probability of receiving a call based on the caller's behavior (arrival time and inter-arrival time) and the reciprocity (number of outgoing calls per incoming call and inter-arrival/departure time), which is given by (20) where $i = 1, 2, 3, \ldots, 24$.

$$P_A(i) = \frac{1}{4}\left[P^a_A(i) + P^b_A(i) + P^c_A(i) + P^d_A(i)\right] \tag{20}$$

Note that the pdf based on arrival time is to be chosen between (3) and (5), depending upon the result of the χ^2-test, that is, a positive result (passing the test) leads to selecting (3) over (5) and vice versa.

There are some callers who never received any calls back from the user, that is, no reciprocity. More likely these callers are telemarketers or voice spammers. Since there

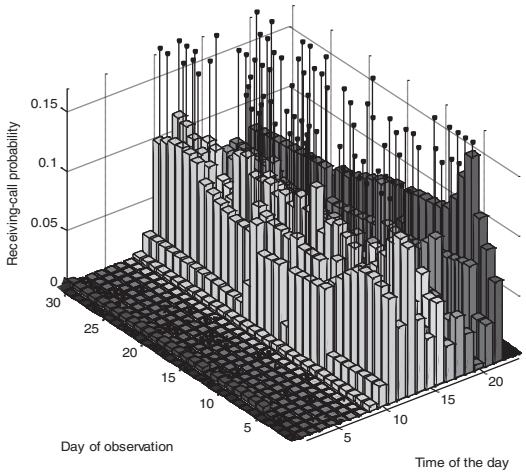

Fig. 7. A randomly selected phone user with 30 consecutive days of computed receiving-call probability of an arbitrary caller, plotted with the actual received calls represented with vertical pulses.

is no history of call interaction between the callers and the user, (20) reduces to the average over the probability based only on the caller's behavior, which is given by (21). Likewise, for the normal callers (not telemarketers or spammers), occasionally some hour slots (rows of call matrix) have no reciprocity, in which (20) also reduces to (21).

$$P_A(i) = \frac{1}{2}\left[P_A^a(i) + P_A^b(i)\right] \tag{21}$$

To present the accuracy of the receiving-call probability model, a phone user is randomly selected from our datasets. Figure 7 shows the computed receiving-call probability of 30 consecutive days for an arbitrary caller where the actual calls during these 30 days are represented with vertical pulses.

From Figure 7, it can be observed that most of the calls are received when the computed receiving-call probability is high and no calls are received during the 0AM to 9AM period, when the probability of receiving a call is low.

4.2. Behavior Trend Detection

So far, we have described the CP which consists of a Probability Estimator (PE) that computes the probability of receiving calls of the next 24 hours based on the call history which is simultaneously collected from the phone user's call activities. However, as the amount of historical call logs increases over time, the computational density can easily overwhelm the PE. Hence the adequacy of historical data has to be identified.

4.2.1. Adequacy of Historical Data. In the previous section, we show that a single-hop caller can be estimated by a normal distribution model $N(\mu, \sigma^2)$, which is characterized by the mean μ and variance σ^2. In attempt to find out how much historical data is actually needed or adequate, we monitor the values of the mean and variance of arrival time for all single-hop callers as more historical data (increased by day) are taken into computations. We observe the convergence of means and variances. As an example, Figure 8 shows the convergence of mean and variance of arrival time of a single-hop caller as the number of days towards the past increases.

It can be observed that the values of mean and variance converge to nearly constant after taking approximately the last 30 days of historical data. This means that the mean and variance of the entire historical call logs are approximately the same as the

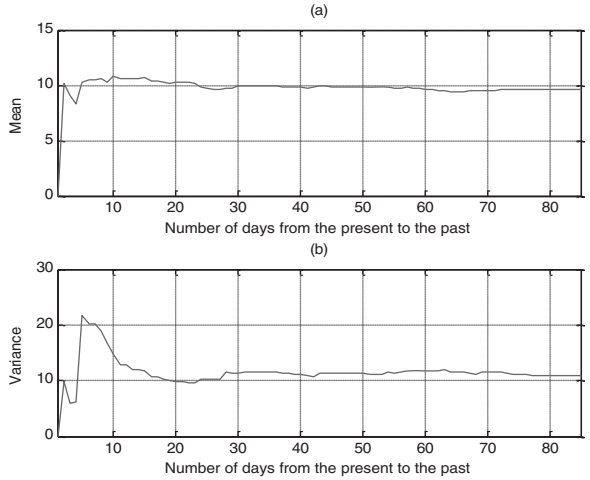

Fig. 8. An example of observed convergence of: (a) mean and (b) variance of arrival time of a single-hop caller.

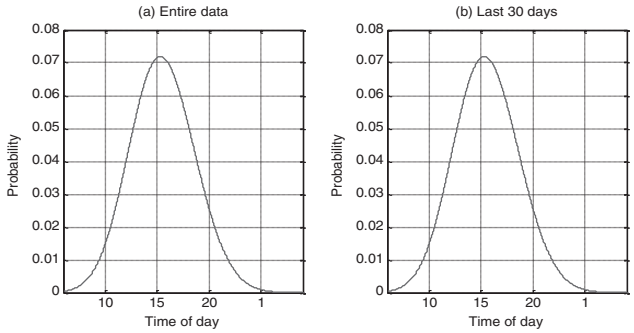

Fig. 9. A comparison of pdf between: (a) taking entire historical data and (b) taking only the last 30 days of call logs.

mean and variance of the last 30 days of call logs. Since a single hop is estimated by a normal distribution, which is characterized by the mean and variance, we can infer that the last 30 days of call logs are adequate to capture the behavior of the single-hop caller. It is evident in Figure 9 that the pdf derived from taking entire historical call logs and taking only the last 30 days are similar.

A knowledge of the mean and variance may not provide enough information to capture the pattern of a multihop caller due to the characteristics of the nonparametric density estimation. However, we believe that it captures physical significance in the behavior of the caller. In fact, the convergence of values of mean and variance of the multihop callers is also observed. As an example, Figure 10 shows that the mean and variance of a multihop caller converge as the number of days towards the past increases.

It is observed that the convergence time for this multihop caller is approximately 60 days. Thus, Figure 11 shows the comparison of the pdf derived from the entire historical call logs and the pdf derived the last 60 days of call logs. The two appear to be slightly different in shape, but the means and variances are nearly the same.

We believe that the call logs represent human behavior associated with trends and changes of behavior over time. Considering historical call logs within the convergence

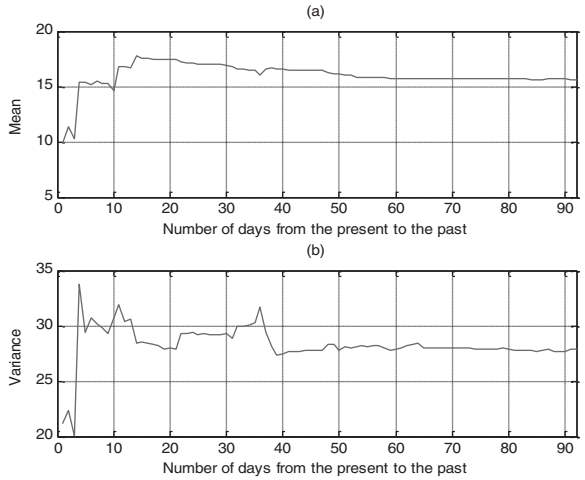

Fig. 10. An example of observed convergence of: (a) mean and (b) variance of arrival time of a multihop caller.

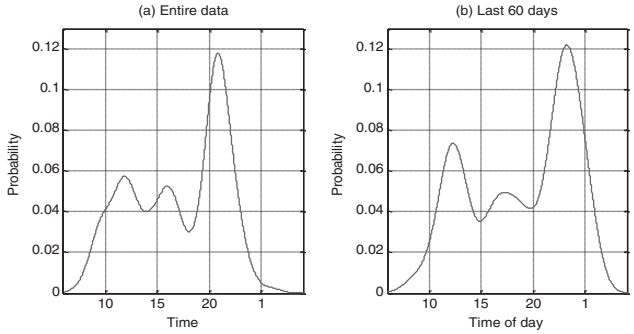

Fig. 11. A comparison of pdf between: (a) taking entire historical data and (b) taking only the last 60 days of call logs.

time may provide us the recent trend of the caller's behavior which can be more relevant to the future observation of behavior.

Our hypothesis is that the future caller's pattern (behavior) is more relevant to the pattern derived from the recent call logs (trend) than the pattern derived from the entire historical call logs (given that the entire call logs are more than the recent trend call logs). This hypothesis will be validated by an experiment conducted later in this section.

We have previously approximated the convergence time for the sample callers shown in Figure 8 and Figure 10 based on visual inspection. To find the exact value of the convergence time, we propose a simple technique for finding convergence time using Trace Distance (tD).

Let us consider a sample of a converging signal shown in Figure 12 where the vertical axis represents the amplitude and the horizontal axis represents reversed time (time that runs towards the past) as similar to the plots shown in Figure 8 and Figure 10.

A trace distance at time $k(tD_k)$ of a signal s is defined as a difference between the minimum and maximum amplitude from time k to infinity (the most right-hand side

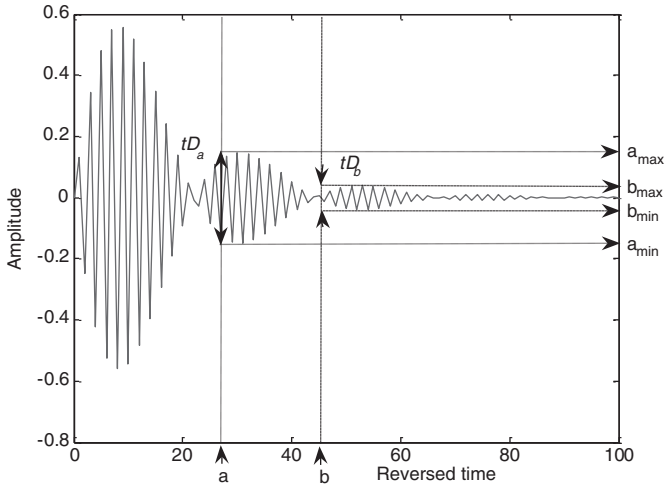

Fig. 12. A converging signal which displays trace distances (tD_a and tD_b) at reversed time a and b for demonstrating convergence time computation.

of time k based on Figure 12) which is given by (22).

$$tD_k = ||k_{\max}| - |k_{\min}|| \tag{22}$$

where k_{max} and k_{min} are defined by (23) and (24), respectively

$$k_{\max} = \max\{s(k), s(k+1), s(k+2), \ldots, s(\infty - 1), s(\infty)\} \tag{23}$$

$$k_{\min} = \min\{s(k), s(k+1), s(k+2), \ldots, s(\infty - 1), s(\infty)\} \tag{24}$$

Thus, the trace distances at time a and b shown in Figure 12 can be computed as $tD_a = ||a_{\max}| - |a_{\min}||$ and $tD_b = ||b_{\max}| - |b_{\min}||$.

Therefore, the Convergence Time (CT) of the signal s is defined as the time that the trace distance (tD) reaches the predefined threshold (tD_{th}) as the trace distance computation starts from reversed time equals to zero to infinity which is given by (25).

$$CT_s = \{k | tD_k = tD_{th}, k \in \{0, 1, 2, \ldots, \infty\}\} \tag{25}$$

For our case, the signal s can be a reversed time series of mean and variance, and the variable k represents the number of days towards the past.

To investigate the relationship of the convergence time between the callers in our datasets, the convergence time is computed for each caller with the tD_{th} set to 1. We find that the convergence time increases as the number of hops increases. Figure 13 shows a plot of the average convergence time versus the number of hops.

We find that the result is reasonable. People who have random behaviors tend not to establish any behavioral pattern in a short period of time, but rather to expand a recognizable structure over longer period of observation time. For example, one caller was initially making several calls in the morning, then started to make some calls in the evening, and eventually making calls consistently in both morning and evening hours (a two-hop caller). It would take more time to observe this caller's calling behavior than another caller who has been calling only during the morning hours (a single-hop caller).

4.2.2. Validation of Hypothesis. To validate our hypothesis that the future caller's pattern (behavior) is more relevant to the pattern derived from the recent call logs (trend) than the pattern derived from the entire historical call logs, an experiment is conducted.

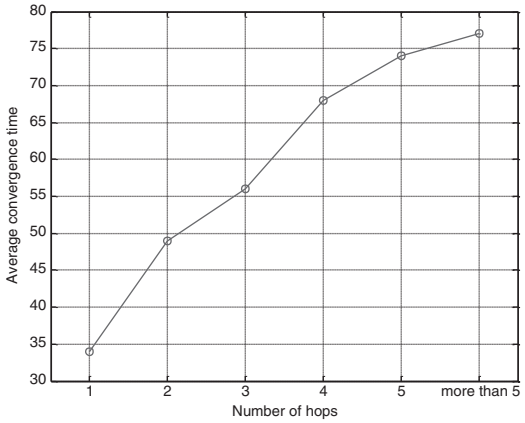

Fig. 13. A plot of the number of hops versus the average convergence time where the average convergence time gets larger as the number of hops increases.

The following experiment is performed to present the comparison of the relevance or similarity in caller's behavior between: (i) the future pattern and the pattern derived from the entire historical call logs, and (ii) the future pattern and the pattern derived from the recent call logs (within the convergence time). To measure the similarity in the calling pattern, the correlation coefficient is chosen.

Correlation coefficient [Leon-Garcia 1994] has value between -1 and 1 which measures the degree to which two random variables are linearly related. A correlation coefficient of 1, -1, and 0 implies perfect linear relationship, inversely proportional relationship, and no linear relationship, respectively. A correlation coefficient (r) can be computed by (26) where P and Q are random variables that consist of small random variables $\{p(1), p(2), p(3), \ldots, p(N)\}$ and $\{q(1), q(2), q(3), \ldots, q(N)\}$ respectively.

$$r = \frac{\sum_{n=1}^{N} (p(n) - \overline{P})(q(n) - \overline{Q})}{\sqrt{\sum_{n=1}^{N} (p(n) - \overline{P})^2 (q(n) - \overline{Q})^2}} \tag{26}$$

In many applications, a correlation coefficient is used to measure how well trends in the predicted values follow trends in past actual values or how well the predicted values from a forecast model fit with the real-life data. In our case, P and Q can be the N-tuple probability mass functions of the future observation and testing period, respectively, where $\overline{P}$ and $\overline{Q}$ are the means of P and Q respectively. Therefore, our testing periods are: (i) entire historical call logs and (ii) within the convergence time.

The experiment is conducted with 100 randomly selected callers including 30 single-hop callers and 70 multihop callers from our datasets. The most recent seven days of the call logs are assumed to be the future observation (pattern). The trace distance threshold (tD_{th}) is set to 1 to compute the Convergence Time (CT).

The results of the computed values of the correlation coefficient between: (i) the future calling pattern and the pattern derived from the entire historical call logs, and (ii) the future calling pattern and the pattern derived from the recent pattern (within the convergence time) are graphically illustrated for comparison in Figure 14(a) where the first 30 callers (callers 1–30) are single-hop callers and the rest of the callers are multihop callers (callers 31–100).

In addition to the comparison purposes, the differences or the changes in the correlation coefficient from considering the entire historical call logs to the convergence time are shown in Figure 14(b).

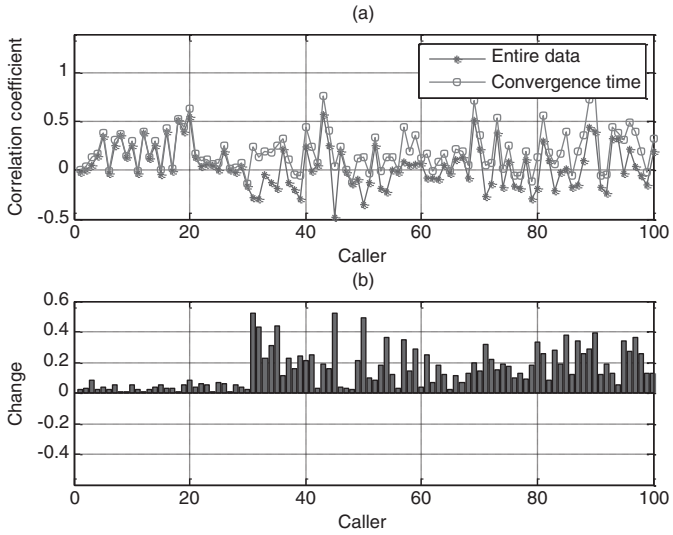

Fig. 14. (a) Comparison of the correlation coefficients and (b) its corresponding change from taking the entire historical call logs to taking call logs within the convergence time of each caller.

Table I. The Average Correlation Coefficient between the Observed Future Pattern and the Pattern Derived from Entire Historical Call Logs compared to the Future Pattern and the Pattern Derived from the Call Logs within the Convergence Time

Callers	Average Correlation Coefficient of Taking Entire Data	Average Correlation Coefficient of Taking Data within Convergence Time	Average Change
Single-hop	0.1280	0.1633	+0.0353
Multi-hop	0.0008	0.2241	+0.2233

It can be observed that the value of the correlation coefficient increases as the convergence time is considered for all 100 callers, which tells us that the recent caller's behavior (pattern) is more relevant (correlated) to the future calling pattern than the pattern derived from the entire call history.

In addition, the experiment is performed for all 4,156 callers among which there are 541 single-hop callers and 3,615 multihop callers. The results are summarized in Table I, which lists average values of the correlation coefficient when the entire historical call logs have been considered as well as when the convergence time has been considered, and their average changes for single-hop and multihop callers.

It can be observed from the graphical representations in Figure 14 as well as the numerical results summarized in Table I that since the single-hop callers have normal distributions (pdf are very similar between deriving from the entire call logs and the convergence time as previously shown in Figure 9), the changes in the similarity measures are relatively low compared to the multihop callers.

Overall, this experimental result shows that the call logs within the convergence time are adequate to capture the caller's calling behavior and in fact they compose a recent trend of the pattern, which is more similar or relevant to the future observed pattern than the pattern composed by the entire historical call logs.

5. OUTGOING CALL PREDICTION FRAMEWORK

To predict the future outgoing calls, similar behavior learning models to those described in Section 4 can be used. Similarly, these models capture the user's behavior (based on

call departure time and interdeparture time), the callee's behavior (based on call arrival time), reciprocity (based on call interarrival/departure time), and user's behavior trend, to construct the probabilistic model for the user's outgoing calls and eventually generate a list of the most likely contacts/numbers to be dialed, which is envisaged as an "Intelligent Address Book."

Similar to the incoming call prediction scenario, as we are at the potential caller's point of view, so we also believe that initiating a call by the user is influenced by the *user's past behavior* (past outgoing calls to the caller) and *reciprocity* (historical call activities between the user and the caller.)

Similar analysis to the incoming call prediction framework can be applied here, where the calling patterns of the user toward different callers can be classified as single-hop and multihop based on the call departure time (i.e., time of initiating calls). Hence the pdf derived from the arrival time of single-hop and multihop callers given by (3) and (5) can be utilized for the departure time for single-hop and multihop calling patterns, respectively.

For a single-hop calling pattern, let a normal random variable R represent the departure time where $R = \{r(1), r(2), r(3), \ldots, r(N)\}$ where N is the total number of outgoing calls and $r(n)$ is the n^{th} call departure time and its pdf is given by (27).

$$e(r) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(r-\mu)^2/2\sigma^2} \tag{27}$$

For a multihop calling pattern, the kernel density estimator is given by (28) where h is the bandwidth and kernel function *K(u)* is given in (6).

$$e(r) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{r - r_i}{h}\right) \tag{28}$$

The behavior of the user can also be characterized by the call inter-departure time. Based on the call matrix shown in Figure 4, the pdf of the interdeparture time can be derived similarly to (10) where variable N is the total number of outgoing calls, $s_k(n)$ is the n^{th} inter-departure time, $S_k = \{s_k(1), s_k(2), s_k(3), \ldots, s_k(N-1)\}$, and $\mu_{S,k}$ and $\sigma^2_{S,k}$ are mean and variance of inter-departure time of the i^{th} hour slot. Its pdf is given by (29).

$$f_i(s_i) = \frac{1}{\sqrt{2\pi\sigma^2_{S,i}}} e^{-(s_i-\mu_{S,i})^2/2\sigma^2_{S,i}} \tag{29}$$

As previously mentioned, initiating a call by the phone user is not only influenced by the user's past calling behavior, but reciprocity as well. Again, similar to the incoming call prediction scenario, the reciprocal calling patterns can be observed via the number of incoming calls per outgoing call which can give us a good time frame approximation of the next outgoing call. With the similar analysis which derives (14), let a random variable U_i represent the number of incoming calls per outgoing call of the i^{th} hour time slot where $\mu_{U,i}$ and $\sigma^2_{U,i}$ are its mean and variance, respectively, hence its pdf can be calculated by (30).

$$g_i(u_i) = \frac{1}{\sqrt{2\pi\sigma^2_{U,i}}} e^{-(u_i-\mu_{U,i})^2/2\sigma^2_{U,i}} \tag{30}$$

Besides monitoring the number of incoming calls per outgoing call, the calling pattern based on reciprocity can also be observed from the inter-departure/arrival time. Similar to the analysis used to derive (18), let random variable V_i represent the

inter-departure/arrival time of the i^{th} hour time slot where $v_i(n)$ is the average inter-departure/arrival time of the n^{th} incoming call to all right-hand-side outgoing calls (across the call matrix's row) prior to reaching the $(n+1)^{th}$ outgoing call. Hence the pdf of the interdeparture/arrival time is given by (31).

$$h_i(v_i) = \frac{1}{\sqrt{2\pi\sigma_{V,i}^2}} e^{-(v_i-\mu_{V,i})^2/2\sigma_{V,i}^2} \tag{31}$$

From (27), (28), (29), (30), and (31), the probability of initiating a call by the phone user to "*Callee B*" at i^{th} hour ($P_B(i)$) can be computed as the average of the probability of initiating a call based on the user's behavior (departure time and inter-departure time) and reciprocity (number of incoming per outgoing calls and inter-departure/arrival time), which is given by (30) where $i = 1, 2, 3, \ldots, 24$ and probabilities ($P_B^a(i)$, $P_B^b(i)$, $P_B^c(i)$, and $P_B^d(i)$) can be calculated in the similar fashion with (4), (11), (15), and (19).

$$P_B(i) = \frac{1}{4}\left[P_B^a(i) + P_B^b(i) + P_B^c(i) + P_B^d(i)\right] \tag{32}$$

Similar to the incoming call prediction scheme, the probability based on departure time is to be chosen between (27) and (28) depending upon the result of the χ^2-test, that is, a positive result (passing the test) leads to selecting (27) over (28) and vice versa.

Again, there is a situation in which some callees never made a single call to the user, hence no reciprocity. Since there is no history of call interaction between the user and the potential callee, (32) reduces to the average over the probability based on only the user's behavior, which is given by (33). Likewise, occasionally some hour slots (rows of call matrix) have no reciprocity, in which (32) also reduces to (33).

$$P_B(i) = \frac{1}{2}\left[P_B^a(i) + P_B^b(i)\right] \tag{33}$$

Again, similar to the incoming call prediction scheme, the same analysis for the behavior trend detection can also be applied here for the user's calling behavior.

6. PERFORMANCE ANALYSIS

In this section, two experiments are conducted to validate the performance of the CP against the actual call logs.

The first experiment is performed to test the performance of the CP as an incoming call predictor, that is, "Incoming Call Forecast." Its performance is measured by the false positives, false negatives, and error rate. A false positive is considered when a call is predicted but no call is received during that hour slot. A false negative is considered when no call is predicted but at least one call is received during that hour slot. The error rate is the percentage of the number of fault predictions to the total number of predictions.

The experiment is conducted with 30 phone users who are randomly selected from our datasets. The call logs of the latest 60 days are assumed to be the future observed call activities where the rest of the call logs (about seven months) represent the call history. For each of the 60 days of testing period, the new call prediction is consequently made by the CP at midnight (0AM) with all available call history (up to that day) taken into account. The trace distance threshold (tD_{th}) is set to 1 to compute the Convergence Time (CT). The computed receiving-call probability is checked with the threshold value to make a decision as to predict "Call" or "No Call" for each of the next 24 hours. The average number of calls per day is computed and rounded to the next largest integer

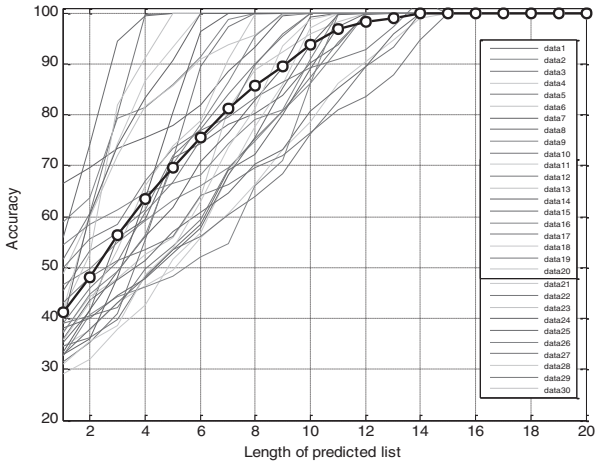

Fig. 15. Performance of the predicted list.

G. The threshold value is dynamically set as *G* hour slots to be selected to make "Call" prediction and the rest of the (24-*G*) hour slots are predicted "No Call." The experimental results are shown in Table A.I (Appendix). Note that the prediction is made for the callers who have at least 20 past incoming calls.

The CP has made a total of 314,784 call predictions with 27,320 fault predictions for these 30 phone users. Note that the number of predictions for each phone user is different due to the different number of associated callers for each user. The average false positive rate is 8.2351%, the average false negative rate is 1.0144%, and the average error rate is 9.2495% (or accuracy rate of about 90.75%). Therefore the average number of fault predictions per day (24 predictions) is 2.2199 with an average tolerance of 2.1407 hours. The average tolerance is a measure of how far off (in hours) the predicted call is from the actual call when the fault prediction occurs.

The second experiment is to test the performance of the CP as an outgoing call predictor, that is, "Intelligent Address Book." The experiment is conducted with the same 30 phone users from the first experiment. Similarly, the latest 60 days of call logs are assumed to be the future observed call activities (testing period) where the rest of the call logs (about seven months) represent the call history (training period). For every call made by the user during the testing period, the CP generates a list of contacts/numbers that the user wants to call and gives these contacts higher precedence on the list. Clearly if the CP performed perfectly, one would expect the actual called number to be at the top of the predicted list. Generally such performance is not achievable, but one might expect that the called number would tend to appear early rather than late in the list.

Table A.II (Appendix) shows the percentage of the actual called numbers that appear within the predicted list as the length of the list varies (1, 3, 5, and 15) for each user. The result shows that on the average if the predicted list is only allowed one entry, the CP correctly predicts the number dialed 41% of the time. If the predicted list has five entries, the CP correctly predicts the dialed number 70% of the time. Finally, if the predicted list contains 15 entries, the dialed number is always present in the list. The average accuracy, which is measured by the percentage of the called numbers in the predicted list, along with the absolute values of all 30 users, is plotted in Figure 15 in black bold line and color lines respectively.

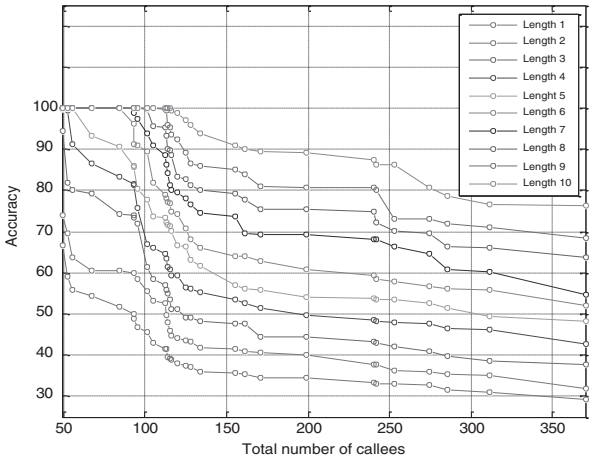

Fig. 16. Accuracy of predicted list at different total number of callees and different length of the list (1-10).

The accuracy of the predicted list is also affected by the population of the called numbers, that is, the number of callees. If the user has been making all calls to only one person, then this user is perfectly predictable. Prediction becomes harder as the number of callees increases (i.e., more possible called numbers). To see this relationship, we plot the total number of callees versus the accuracy of the CP for 10 different lengths of the list in Figure 16.

7. ANALYSIS OF AUTONOMY AND ADAPTATION

As the autonomy and adaptation of computing systems are the subjects of interest of ACM TAAS readers, it is thus important to reemphasize these properties of the proposed model of CP as well as present the useful applications of CP by which the smart phone becomes smarter with these properties—more adaptive and responsive. Therefore this section aims to describe autonomous and adaptive characteristic aspects of the CP as well as its applications.

7.1. Autonomous and Adaptive Characteristics of CP

The CP exploits the historical call logs to construct a probabilistic calling behavior model. Human behavior tends to repeat periodically and creates a pattern that changes over time. Since call logs are also human behavioral data, it follows the same characteristic. In fact, recent behavior is more relevant to future behavior than the old one. Thereby the recent trend of behavior is detected with TD. The call prediction can then be made based on the recent behavior calling pattern either by the user's request (as in incoming call prediction scheme) or automatically (as in outgoing call prediction scheme). Autonomy and adaptation of CP are evidenced by the recent behavior detection mechanism that allows the model to adapt its computation in response to the new calling behaviors. Hence the prediction is made adaptive to the most recent behavior. Moreover, this mechanism provides a means to automatically remove unnecessary data such that the model remains computationally feasible as more data arrive. Figure 17 shows the architecture of CP displaying a feedback loop for updating stored call logs so that the CP remains adaptive to the user/caller's recent calling behavior.

To demonstrate the impact of the feedback control, we conduct an additional experiment by testing CP with different update feedback rates and recording their corresponding error/accuracy rates. The experiment is conducted with 30 users and the latest 60 days are used for testing while the rest are used as training data. Figure 18

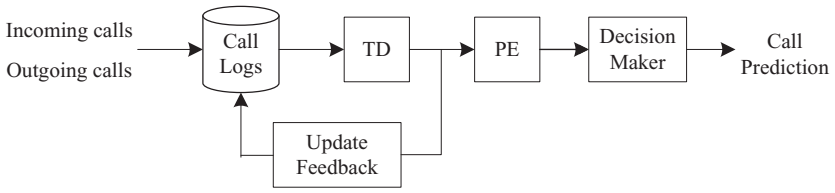

Fig. 17. Architecture of CP showing feedback loop such that it continues to be adaptive to the user/caller's recent behavior.

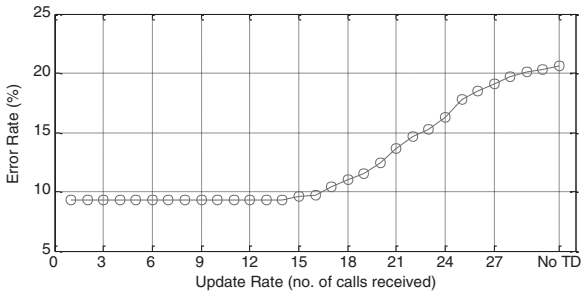

Fig. 18. The update rate (number of calls received) and its corresponding error rate of CP as an incoming call predictor. The last reading (farthest right) is the error rate of CP without using TD.

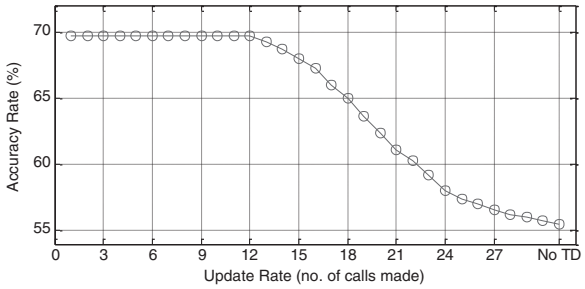

Fig. 19. The update rate (number of calls made) and its corresponding accuracy rate of CP as an outgoing call predictor. The last reading (farthest right) is the error rate of CP without using TD.

shows the result of CP used as an incoming call predictor while Figure 19 shows the result of CP as an outgoing call predictor (at list length = 5). The error rate of the incoming call predictor starts from the minimum value at 9.25% when the update is done for every incoming call and it remains at this value until the update rate is 14 when it begins to increase. Figure 18 also shows that without TD, the error rate would have been 20.60%. This result tells us two things. First, the feedback control is essential for the model to remain adaptive as evidenced from the increase of error rate of 11.35% by not using TD (that is, no feedback control). Second, the update rate does not have to be at every incoming call; based on the result from this testing dataset, the update rate can be set at every 14 calls such that it saves computational cost as it usually means power. Similar observations can be made for the result of the outgoing call predictor; the accuracy rate starts off at a constant maximum rate then begins to drop at the update rate of 13 outgoing calls, and a gap of 14.18% difference in accuracy rate can be observed between the model with and without TD.

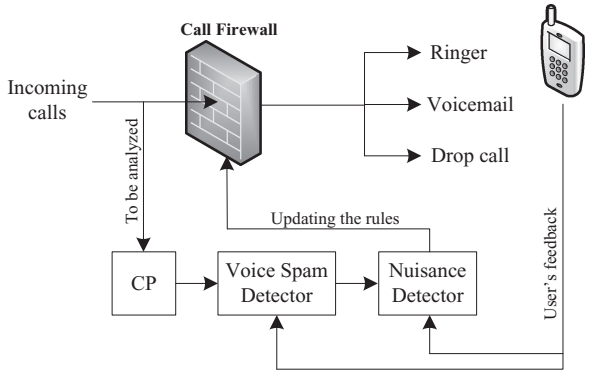

Fig. 20. System overview of Call Firewall constructed with CP, VSD, and ND for proactively handling the incoming calls.

7.2. Applications of CP

To demonstrate the usefulness of CP besides its own features; Incoming Call Forecast and Intelligent Address Book, we describe here two applications of CP including Call Firewall and Call Reminder.

7.2.1. Call Firewall. By adopting the concept of firewall (the wall that keeps destructive forces away from our computer systems), Call Firewall basically monitors and handles incoming calls by keeping unsolicited and unwanted calls away while allowing desired calls to pass through. The problem of unwanted telemarketing calls or spam calls is expected to be a serious problem especially in VoIP networks due to its much lower communication cost than the circuit-switched telephone network system (it also becomes an attractive target for spammers). In fact, SPIT (Spam over Internet Telephony) is roughly three orders of magnitude less expensive to generate than traditional circuit-based telemarketing calls [Rosenberg et al. 2006]. Unlike email spam, call spam is a real-time problem which requires a real-time defense mechanism. The real challenge is thus to block the spam call before the phone rings. Not only do these spam calls create a nuisance for the user, Kolan and Dantu [2008] showed that each incoming phone call creates different levels of nuisance depending on the user's presence (mood or state of mind) based on situational, spatial, and temporal contexts. Therefore, to address this problem of unwanted calls, the system for detecting voice spam and estimating spamminess level (known as VoIP Spam Detector or VSD) described by Kolan and Dantu [2007] and Dantu and Kolan [2005] and the nuisance computation model (known as Nuisance Detector or ND) proposed by Kolan and Dantu [2008] can be integrated with the call prediction model proposed in this article (CP) to proactively handle incoming calls before the phone rings. VSD, as described in Kolan and Dantu [2007] and Dantu and Kolan [2005], is a multistage adaptive spam filer based on presence (location, mood, time), trust, and reputation to spam in voice calls. It uses a close-loop feedback control between different stages to detect a spam call. As described in Kolan and Dantu [2008], ND is a model for computing the nuisance level of incoming calls based on the social closeness and other behavioral patterns such as periodicity of the caller and reciprocity.

As shown in Figure 20, CP generates a periodic 24-hour call prediction to be fed into VSD to learn the behavior of callers (among which are spammers) and analyze trustworthiness (VSD indicates the untrusted calls to be "dropped") and ND computes the nuisance level associated with each predicted call (ND determines each call to be either sent directly to "voicemail" or "ringer" to ring the phone), then a set of firewall

rules is generated, for example, IF John calls between 10am–11am, THEN forward it to voicemail, IF Pizza House calls between 4pm–5pm, THEN drop the call. The firewall rules are updated periodically (can be as often as every hour, depending on the user). The user can also provide feedback about the actual nuisance level or reporting spam calls in order to improve the performance of the firewall. In summary, the proposed Call Firewall is an autonomous system that is adaptive to recent calling behaviors as it proactively manages incoming calls based on the preconfigured set of rules by keeping unsolicited calls away while allowing wanted calls to either ring the phone or be forwarded to voice mail (if nuisance level is high).

To show the performance of the Call Firewall, an experiment is conducted with 30 users (latest 60 days are testing data while the rest are training data). Table A.III (Appendix) shows the false negative rate, true negative rate, and true positive rate of all 30 users. False negative rate measures the percentage of the incoming calls that pass through the Call Firewall but should have been blocked. We assume that all "Missed Call" in our dataset means that the user does not want to take the call and hence it should be blocked by the Call Firewall. Despite many other reasons for the missed calls such as being away from the phone, not hearing the ringer, and forgetting to switch the phone back to ringer from silent mode, we carry out the experiment with this assumption. True negative rate is a percentage of correctly blocked calls by Call Firewall, that is, (number of blocked calls)/(number of predicted calls to be blocked). True positive rate is a percentage of calls that are correctly let through by Call Firewall, that is, (number of pass-through calls)/(number of predicted calls to be allowed to pass through firewall). Based on this experiment, the Call Firewall performs with the average false negative rate of 10.40%, true negative rate of 75.70%, and true positive rate of 83.03%.

7.2.2. Call Reminder. One of the common problems of everyday life is forgetting to make a phone call that could either be an event-based call such as birthday call, meeting planning call, etc., or a nonevent-based call such as calling parents on weekends, calling girlfriend/boyfriend during a lunch break, etc. Therefore, besides the Intelligent Address Book (an automatic function that computes the probability of outgoing calls based on recent calling behavior and generates a list of potential callees to help avoid searching for a number to call through a typical lengthy address/contact book) we present here a Call Reminder that makes use of CP as an outgoing call predictor by integrating it with ND and an Event Calendar to generate a "reminder" for the user to place a call to a particular person based on the user's past history, nuisance level, and events.

As shown in Figure 19, CP periodically makes outgoing call prediction (e.g., hourly), which will be mapped onto the nuisance level computed by ND. The result is then evaluated by the decision maker to generate the call reminder, for example, high probability and low nuisance level would imply prompting a call reminder. The event calendar (a function that normally comes with today's mobile phones) is used to provide details about the call reminder, for example, birthday call, meeting plan, project discussion, etc. The user would be prompted with a reminding message such as "Would like to call John about the ABC conference?", "Would like to call Alice about the birthday?", "Would you like to call Mom regarding dinner?". The user records new events into the event calendar for future reminders. Feedback sensor forwards the actual outgoing calls to CP to be analyzed for prediction as well as provides the user's feedback to ND to calibrate nuisance computation. In summary, the proposed Call Reminder is an autonomous system that is adaptive to the recent calling behaviors as it provides an automatic reminder for placing a call based on the probability of making a call to a particular person, nuisance level of the user, and associated events (Figure 21).

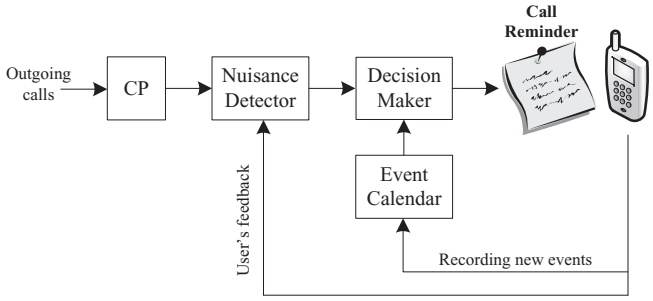

Fig. 21. System overview of Call Reminder constructed with CP, ND, and Event Calendar for reminding the user to place a call.

To see the performance of the Call Reminder, we conduct an experiment with call logs of 30 users where the latest 60 days comprise the testing period while the rest of the data are used for initial training. Since there is no event calendar information in our dataset, the performance is based solely on CP and ND. The goal is to measure the percentage of the calls made because of the prompted reminders generated by the Call Reminder. Our assumption here is that each outgoing call needs to be reminded. Clearly, it is not completely realistic. Nonetheless, to get the first glance at how Call Reminder would perform in real life, we conduct the experiment on this assumption. Therefore, with our dataset, we verify for each outgoing call if it would be reminded by the Call Reminder. Table AIV (Appendix) shows the result of the true positive rate, which is computed as ratio of the number of actual outgoing calls made that are among the five numbers/contacts reminded by the Call Reminder to the total number of outgoing calls. Based on this experiment, the Call Reminder performs with the average true positive rate of 69.27%. Note that we believe that the performance can be relatively improved with an event calendar. Moreover, in more realistic setup, only some outgoing calls should be reminded. To see the real performance of the proposed Call Reminder, one would be interested in finding out if the user does make an outgoing call when a call reminder is generated. This will be in our future study.

8. RELATED WORK

There have been some literatures on predictive models for telephone call demands. In Brown et al. [2005], the authors applied queuing theory to characterize queuing primitives such as the arrival time process, the service-time distribution, and the distribution of customer impatience. In Aldor-Noiman [2006], the author developed two variations of Poisson process models for describing count data of call center arrivals that utilized the proposed mixed models technique. There is also literature describing a predictive model for the emergency 9-1-1 call volumes ([Jasso et al. 2007]) where the authors used a multiple linear regression model technique to construct the proposed multidimensional linear predictor based on call history.

To the best of our knowledge, the literature that is closest to our work is Harless and Kowalski [2000]. The authors developed a system for predicting a future communication activity based on the past communication event information. The system analyzed the past communication event information (including phone calls and emails) to determine whether a correlation existed in the past communication and predicted the future communication event based on the current communication event and the correlation. The correlation is computed based on the pattern of incoming and outgoing calls, for example, if a call received from "person A" results in a later origination of a call to "person B," the correlation value between the "person A" and the

"person *B*" is increased proportionately and the correlation values corresponding to other persons not dialed is decreased accordingly.

In contrast, our work is focused to predicting the future incoming calls for the next 24 hours and outgoing calls in form of a list of the most likely contacts/numbers to be dialed based on the past communication information of the phone user and the potential callers/callees. We do not correlate the callers using their past communication patterns towards the phone user. However, we think that considering the temporal sequence of call activities in computing the probability of receiving/initiating a future call may improve the accuracy of the CP for which will be discussed further in our future work.

9. CONCLUSION

Over the past few years, there has been a rapid development and deployment of new advanced phone features, including Internet access, email access, scheduling software, built-in cameras, contact management, accelerometers, and navigation software, as well as the ability to read documents in variety of formats such as PDF and Microsoft Office. However, none of these features offers the ability to predict future calls. In this article, we propose a Call Predictor (CP) that can operate as an incoming call predictor, which offers a new feature for the next-generation phones known as "Incoming Call Forecast," and an outgoing call predictor, which provides an improvement over the "last numbers dialed" functionality that is often provided on the phones by an "Intelligent Address Book."

As an incoming call predictor, the CP computes receiving-call probability based on the caller's past behavior and the reciprocity, and makes the next-24-hour call prediction. The caller's past behavior is characterized by the past incoming call pattern, which can be observed by the arrival time and inter-arrival time. On the other hand, the reciprocity is characterized by the number of outgoing calls per an incoming call pattern, and the inter-arrival/departure time pattern. We believe that call logs include trends or changes of human behavior over time. In fact, we have proved that the recent trend of the caller's behavior has higher correlation to future behavior than the patterns derived from the entire historical data. Thus, the CP detects the recent trend of the caller's behavior and computes the adequacy of historical data in terms of the reversed time based on the trace distance to reduce the computational density.

As an outgoing call predictor, the CP computes the probability of initiating a call to each callee based on the user's past calling behavior towards each callee and reciprocity, and it generates a list of the most likely contacts/numbers to be dialed at a given time. The user's past behavior is characterized by the past outgoing call pattern which can be observed via the call departure time and inter-departure time. The reciprocity or the past call interaction activity is characterized by the pattern of the number of incoming calls per outgoing call and the inter-departure/arrival time. Similar to the incoming call prediction scheme, the CP also detects the recent trend of the user's calling pattern and computes the adequacy of historical call logs as the recent calling pattern is proven more relevant to future observation than the pattern derived from the entire historical call logs.

The performance of the CP is validated against the actual call logs of two months based on the historical call logs of seven months. The result of the Incoming Call Forecast shows a fairly good performance with low false positives, false negatives, and error rate. Likewise, the Intelligent Address Book also shows a promising result of its performance. Nevertheless, the prediction technique proposed here is preliminary and other approaches need to be considered in order to improve the performance of the predictor. In addition, we describe two applications of CP including Call Firewall and Call Reminder. Call Firewall proactively manages incoming calls based on preconfigured set of rules by keeping unsolicited calls away while allowing wanted calls to either

ring the phone or be forwarded to voice mail. Call Reminder provides an automatic reminder for placing a call based on the probability of making a call to a particular person, nuisance level of the user, and associated events. Our future work will involve in investigating other parameters to characterize the behavior of the phone users and examining other prediction techniques to improve the performance of the CP.

APPENDIX

Table A.I. The Experimental Results for Validating the Performance of the CP as an Incoming Call Predictor

Phone user	Number of predictions	Number of fault predictions	False positive (%)	False negative (%)	Error rate	Number of fault predictions per day	Average tolerance (hours)
1	7440	746	9.4515	0.5754	10.0269	2.4065	1.7379
2	5952	748	11.7444	0.8228	12.5672	3.0161	1.7621
3	7320	687	9.0113	0.3739	9.3852	2.2525	1.5182
4	16728	1353	7.4419	0.6463	8.0882	1.9412	2.3943
5	9216	665	6.6384	0.5773	7.2157	1.7318	1.6243
6	16992	1902	10.552	0.6415	11.1935	2.6864	2.8270
7	24408	2391	9.3439	0.4521	9.7960	2.3510	2.0253
8	4320	246	5.6481	0.0463	5.6944	1.3667	2.2054
9	14544	1191	7.5894	0.5995	8.1889	1.9653	1.9231
10	6072	787	9.8646	3.0965	12.9611	3.1107	2.2347
11	12744	1439	10.2307	1.0609	11.2916	2.7100	1.8578
12	2208	259	8.5532	3.1769	11.7301	2.8152	2.1824
13	8472	766	8.1713	0.8702	9.0415	2.1700	1.5090
14	10992	1396	11.3392	1.3609	12.7001	3.0480	2.1659
15	25296	1644	4.7204	1.7787	6.4991	1.5598	1.9186
16	23400	1117	3.8552	0.9183	4.7735	1.1456	2.3365
17	6864	291	2.6285	1.6110	4.2395	1.0175	2.3299
18	5688	301	3.5764	1.7154	5.2918	1.2700	2.4003
19	11472	961	7.9437	0.4332	8.3769	2.0105	2.1440
20	2880	344	11.3194	0.6250	11.9444	2.8667	1.9922
21	6312	724	9.7778	1.6924	11.4702	2.7529	2.1722
22	11592	1210	9.6047	0.8335	10.4382	2.5052	1.8449
23	12048	1355	10.5679	0.6788	11.2467	2.6992	1.7855
24	21240	1325	5.2053	1.0329	6.2382	1.4972	2.1254
25	2880	281	8.9931	0.7638	9.7569	2.3417	2.3516
26	8640	857	9.3634	0.5556	9.9190	2.3806	2.9585
27	6144	634	9.4502	0.8688	10.3190	2.4766	2.4200
28	4656	567	10.735	1.4428	12.1778	2.9227	2.7085
29	15144	841	4.8884	0.6650	5.5534	1.3328	2.3532
30	3120	292	8.8426	0.5164	9.3590	2.2462	2.4136

Table A.II. The Experimental Results of the CP as an Outgoing Call Predictor

Phone user	Total number of outgoing calls	Percentage of called numbers in the predicted list				
		List length = 1	List length = 3	List length = 5	List length = 10	List length = 15
1	865	35.2959	55.8273	71.5278	90.1269	100
2	303	58.9858	79.2705	86.0320	100	100
3	155	51.7157	80.1469	100	100	100
4	625	34.5789	43.0000	49.5263	78.5789	100
5	174	39.0151	42.0454	54.1667	87.5000	100
6	822	37.2976	47.7584	55.8530	95.8281	100

Table A.II. Continued

7	887	66.5983	73.4289	77.8005	100	100
8	409	55.9911	94.4526	100	100	100
9	694	46.7278	54.8430	63.2199	86.2566	100
10	898	31.4368	40.8621	71.2068	100	100
11	529	29.1487	37.7694	51.5624	97.0366	100
12	517	45.4698	56.8792	72.0918	100	100
13	1203	48.8699	71.9326	85.6780	100	100
14	489	33.3933	49.2206	61.6907	99.3405	100
15	947	36.0040	47.5731	57.1475	80.8177	100
16	592	54.3795	61.3139	66.4233	90.8759	100
17	543	34.5699	39.7312	53.4946	76.2903	100
18	796	41.6348	74.3296	93.2312	100	100
19	654	49.9174	58.4983	73.3498	89.3564	100
20	290	41.6071	81.9643	100	100	100
21	132	39.5833	73.9583	90.6250	100	100
22	576	32.9060	43.1623	53.4188	89.1025	100
23	705	32.7329	44.3321	48.3744	86.1599	100
24	609	43.0625	51.3125	56.0626	76.5626	100
25	158	30.9160	38.5496	53.8168	98.8549	100
26	368	39.1304	51.0869	66.6667	100	100
27	337	35.7744	53.6195	73.8215	100	100
28	200	32.9268	48.1707	80.4878	100	100
29	278	38.1278	44.5205	52.7398	93.8356	100
30	378	37.2659	49.2510	70.2248	100	100

Table A.III. The Experimental Result of the Performance of the Call Firewall

Phone user	False negative (%)	True negative (%)	True positive (%)
1	7.9167	60.5657	81.1345
2	13.6364	70.6599	78.7067
3	20.1220	68.9306	85.9894
4	18.5185	73.4268	84.3915
5	2.8571	71.2957	73.0827
6	6.7805	90.5238	89.2000
7	30.6250	70.9195	68.8482
8	29.0909	73.9037	74.4131
9	5.1049	62.6718	80.4147
10	17.0000	61.2121	74.1722
11	0	85.5233	83.5608
12	26.4151	75.4762	86.7951
13	0	95.4762	100
14	3.3333	74.9153	85.2941
15	3.6364	69.6364	74.5827
16	2.5641	70.0632	81.7840
17	6.6667	61.3043	70.0880
18	30.1887	79.5116	77.1222
19	21.5686	79.4787	80.9110
20	4.5455	68.6689	90.8333
21	0	77.4271	84.6699
22	2.1739	69.8182	82.2090
23	12.1951	75.6426	78.1295
24	0	93.5915	94.1000
25	7.1429	86.4356	92.8571
26	0	71.7921	94.5113
27	8.4337	86.1017	86.8067
28	18.3019	76.2393	82.7376
29	11.9048	88.4960	85.3029
30	1.2195	81.2656	88.0456

Table A.IV. The Experimental Result of the Performance of the Call Reminder

Phone user	True positive (%)	Phone user	True positive (%)
1	70.5202	16	65.8784
2	86.1386	17	53.0387
3	100	18	92.0854
4	49.6000	19	72.9358
5	54.0230	20	100
6	54.7445	21	90.9091
7	76.3247	22	52.0833
8	98.7775	23	47.9433
9	62.3919	24	55.9934
10	70.7127	25	53.7975
11	51.0397	26	66.5761
12	72.1470	27	73.8872
13	85.4530	28	80.5000
14	61.5542	29	52.5180
15	56.8110	30	69.5767

REFERENCES

ALDOR-NOIMAN, S. 2006. Forecasting demand for a telephone call center: Analysis of desired versus attainable precision. Master thesis 2005. Department of Statistics, Technion—Israel Institute of Technology.

BROWN, L., GANS, N., MANDELBAUM, A., SAKOV, A., SHEN, H., ZELTYN, S., AND ZHAO, L. 2005. Statistical analysis of a telephone call center: A queeing-science perspective. *J. Amer. Stat. Ass. 100*, 469, 36–50.

DANTU, R. AND KOLAN, P. 2005. Detecting spam in VoIP networks. In *Proceedings of USENIX, SRUTI (Steps for ReducingUnwanted Traffic on the Internet)*.

EAGLE, N. AND PENTLAND, A. 2005. Social serendipity: Mobilizing social software. *IEEE Pervas. Comput. 4*, 2.

EAGLE, N. AND PENTLAND, A. 2006. Reality mining; Sensing complex social systems. *Personal Ubiquit. Comput. 10*, 4.

EAGLE, N., PENTLAND, A., AND LAZER, D. 2007. Infering social network structure using mobile phone data. *Proc. Nat. Acad. Sci.* To appear.

HARLESS, C. E. AND KOWALSKI, T. J. 2000. U.S. Patent 6084954, Jul. 4.

JASSO, H., FOUNTAIN, T., BARU, C., HODGKISS, W., REICH, D., AND WARNER, K. 2007. Prediction of 9-1-1 call volumes for emergency event detection. In *Proceedings of the 8th Annual International Digital Government Research Conference*, vol. 228. 148–154.

JONES, M., MARRON, J. S., AND SHEATHER, S. J. 1996. A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc. 433*, 91, 401–407.

KOLAN, P. AND DANTU, R. 2007. Socio-Technical defense against voice spamming. *ACM Trans. Auton. Adapt. Syst. 2*, 1.

KOLAN, P. AND DANTU, R. 2008. Nuisance level of a voice call. *ACM Trans. Multimedia Comput. Comm. Appl. 5*, 1.

LEON-GARCIA, A. 1994. *Probability and Random Processes for Electrical Engineering*. 2nd Ed. Addison-Wesley.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY. 2007. Reality minning. http://reality.media.mit.edu

NTULI, D. 2007. Phone dating takes off in mobile crazy SA. http://mybroadband.co.za/news/Cellular/1019.htm

PARZEN, E. 1962. On estimation of a probability density function and mode. *Ann. Math. Statist. 33*, 3.

ROSENBERG, J., JENNINGS, C., AND PATERSON, J. 2006. The session initiation protocol (SIP) and spam. Spam Draft-draft-ietfsipping-spam-02.txt

SHEATHER, S. J. AND JONES, M. C. 1991. A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. B, 53*, 683–690.

TRENDHUNTER. 2006. Shopping by phone takes off. http://www.trendhunter.com/trends/shopping-by-phone-takes-off/

WAND, M. P. AND JONES, M. C. 1994. Multivariate plug-in bandwidth selection. *Comput. Statist. 9*, 97–117.

WEB JAPAN. 2006. Reading on the move. http://web-japan.org/trends/business/bus061211.html

Received September 2009; accepted August 2010